Executive Summary

Carvana is an online used car dealership in the United States that buys and sells used cars through your devices or their famous car vending machines. For this project, our team at Hawkeye Consulting collaborated with Carvana to help set the prices for their used car inventory.

Our goal is to determine what factors of a used car drive their prices and whether or not Carvana should classify whether it is expensive or cheap. Additionally, they would be able to attract loyal customers. The use of our model allows Carvana to show customers what features in the car are significantly determining its price.

The used car data used to create our models contains 2000 instances of 10 categorical and 6 numeric features. Our data was cleaned by removing nonsensical data and adjusting formatting errors.

We created three models, decision tree, random forest, and SVM. With a goal of predicting whether or not a car would be identified as cheap or expensive. Based on our findings we were able to determine which factors are affecting a car's price. Some of the factors we found to be affecting the price the most were airbags, right hand drive, engine volume, and the type of gearbox that a car has.

Finally, we will conclude our report with a recap of our dataset and the problem we faced. Also, discussing key recommendations for Carvana including focusing pricing and marketing strategies around safety, interior quality, and fuel size. From there, we will be able to analyze the dataset limitations it imposes. These limitations on our dataset include not having information on technological features, miles per gallon of each car, and whether or not the car has a back-up camera. This allows us to have a better understanding of our dataset and the important factors we have found to fully learn the importance of what makes a cars price.

Background

At Hawkeye Consulting we aim to give solutions to our clients through data visualization and driven recommendations. Therefore, we partnered with Carvana, a used car dealership. Carvana is currently the fastest growing online used car dealership in the US and is mainly known for their car vending machines. Additionally, according to CNBC, American consumers have been paying $10,000 more on average for used cars. The Covid-19 pandemic and the current microchip shortage help explain this. The microchip shortage is causing a new car shortage due to the fact that microchips are an essential part of manufacturing. When consumers physically cannot buy new cars, they are forced to choose from the used car market, one that is known to overprice their products. Mileage is the first thing that comes to mind when thinking about used car prices, however we wanted to determine what other factors contribute to the price of a used car.

Business Goal and Data Mining Goal

Our business goal for Carvana is to determine what factors contribute to the price of a used car and to predict whether it should be priced over or under $15,000 dollars. Determining features that affect the price and classifying whether it should be expensive or cheap will help Carvana attract customers by setting transparent prices and attracting loyal customers. Customer loyalty in car dealerships, especially used, is hard to increase. However, we believe that treating customers fairly in their car shopping experience will help attract loyal customers. Our data mining problem is classification, and our prediction target is whether a car should be labeled “expensive” or “cheap”.

Data Description

Raw Data

ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color
45654403	13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005 km	6	Automatic	4x4	4-May	Left wheel	Silver
44731507	16621	1018	CHEVROLET	Equinox	2011	Jeep	No	Petrol	3	192000 km	6	Tiptronic	4x4	4-May	Left wheel	Black
45774419	8467	-	HONDA	FIT	2006	Hatchback	No	Petrol	1.3	200000 km	4	Variator	Front	4-May	Right-hand d	Black
45769185	3607	862	FORD	Escape	2011	Jeep	Yes	Hybrid	2.5	168966 km	4	Automatic	4x4	4-May	Left wheel	White

Cleaned Data

ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color
45801743	expensive	-	TOYOTA	Corolla	2014	Sedan	No	Petrol	1.8	40000 km	4	Automatic	Front	4	Left wheel	Green
44452971	expensive	1399	TOYOTA	Sienna	2010	Minivan	Yes	Petrol	3.5	227200 km	6	Tiptronic	Front	4	Left wheel	White
45805890	expensive	1055	LEXUS	RX 350	2013	Jeep	Yes	Petrol	3.5	73000 km	6	Tiptronic	4x4	4	Left wheel	Black
45612642	cheap	986	HYUNDAI	Sonata	2010	Sedan	No	Petrol	2.4	90000 km	4	Automatic	Front	4	Left wheel	Red

Data

The raw dataset is from Kaggle.com, is named the Car Price Prediction Challenge, contains 19237 instances, and 18 features. The original dataset includes features such as manufacturer, model, production year, mileage, color, etc. After the data was cleaned, it contained 2000 instances with 10 categorical features and 6 numeric features. Price is our target variable.

Data Pre-processing

- Step 1: Removed nonsensical data.
    - Nonsensical data : Removed unrealistic price data i.e., cars made later than 2010 with a listed price under $1000. This assumes no cars in the dataset needed significant repair.
- Step 2: Corrected data format.
    - Format correction: The ‘doors’ column was formatted as a date. Changed format to integer.
- Step 3: Decreased the number of observations to 2000.
- Step 4: Changed target variable to categorical.
    - Target variable: Changed all values in the ‘price’ column to cheap or expensive. Values greater than or equal to the median price, $15,000, were changed to expensive. Values less than the median price, $15,000, were changed to cheap.

Data Mining Solution

Through various models we were able to find various attributes in a car that affected whether or not the price would go up or down. Airbags, right-hand drive, and levy were the top three most important variables affecting price in our model. Carvana does not currently have all of these things listed in their description of their cars and could potentially be losing or missing sales because of it.

Models

We tested three separate models in order to find the best predictor for whether or not a car would be placed in the expensive or cheap category.

- Decision Tree
- Random Forest
- SVM

The decision tree model provided the error matrix shown below:

	Predicted		
Actual	Cheap	Expensive	Error
Cheap	85	71	45.5
Expensive	29	115	20.1

An accuracy of 67%, precision of 62%, and recall of 80% were calculated on the above error matrix. However, in order to compare all three models, we scored each of them based upon their AUC. The decision tree model was the worst of the tree models with an AUC of 0.71. SVM was the second-best model with an AUC of 0.79, leaving our best model as the random forest with an AUC of 0.85.
As our best model, we used the random forest to determine which attributes of the car were affecting whether or not it would be identified as cheap or expensive. The top five results for variable importance are displayed in below table along with their Mean Decrease Accuracy.

Attribute	**Mean Decrease Accuracy**
Airbags	30.63
Right hand drive	28.26
Levy	26.43

Engine Volume	25.65
Gear-box type (Tiptronic)	25.57

Using these results, we were able to determine which attributes would assist in the fair and competitive pricing for Carvana.

Conclusion

To conclude, we discussed how a high percentage of cars are being sold at a premium. From there, we set our business goal to see what additional factors can impact a car's price. Using a Random Forest, Decision-Tree and SVM model we were able to interpret our dataset and find factors with importance to form our overall recommendations and limitations to our set.

Recommendations

After creating a Random Forest, Decision-Tree, and SVM model based on our findings, we recommend that Carvana focus their purchasing and marketing strategy around safety, interior quality, fuel size, and size class. Based on our dataset, we found that car safety has a higher mean decrease accuracy when performing a random forest model. This helped form an overall recommendation based around safety to adequately show the variable importance represented in our random forest model. The same can be said for fuel size because of the importance it has when customers purchase a car. Also, interior quality and size class are recommended when focusing on purchasing and marketing strategies because of the appeal it can have on customers. This can lead to higher increases in revenue and marketing appeal.

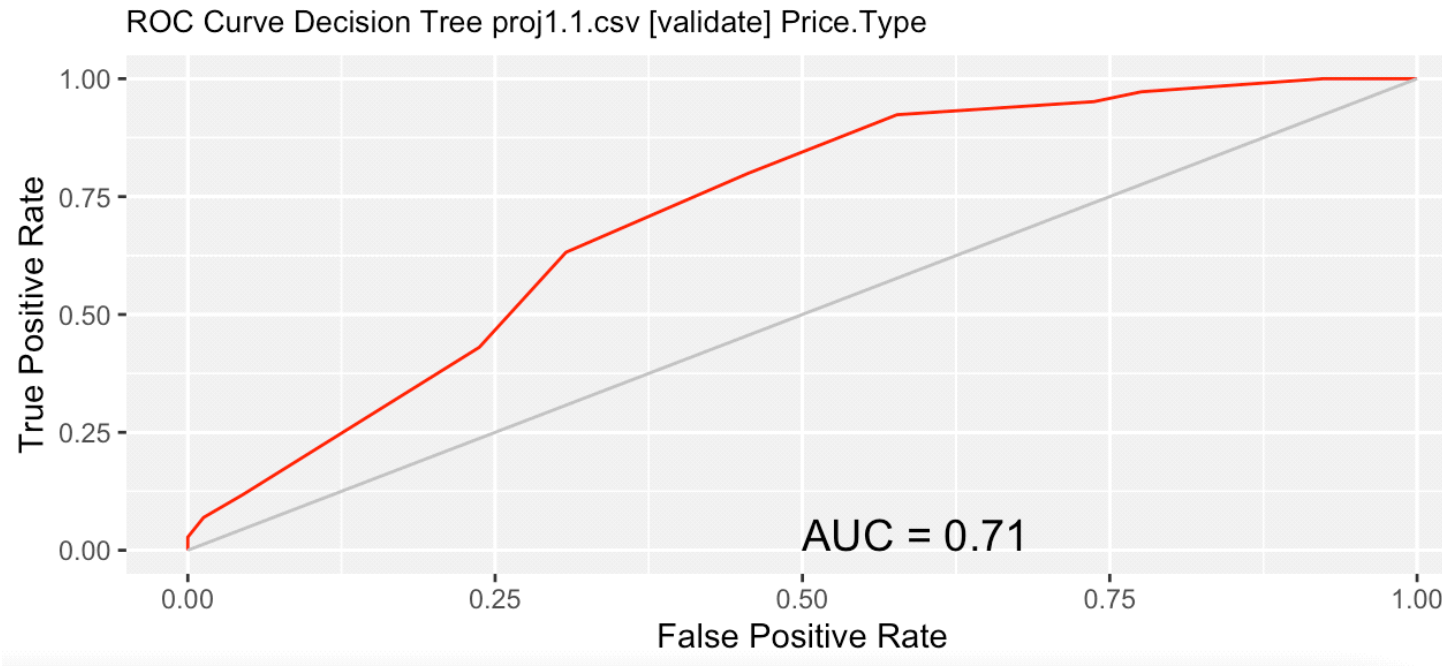
Limitations

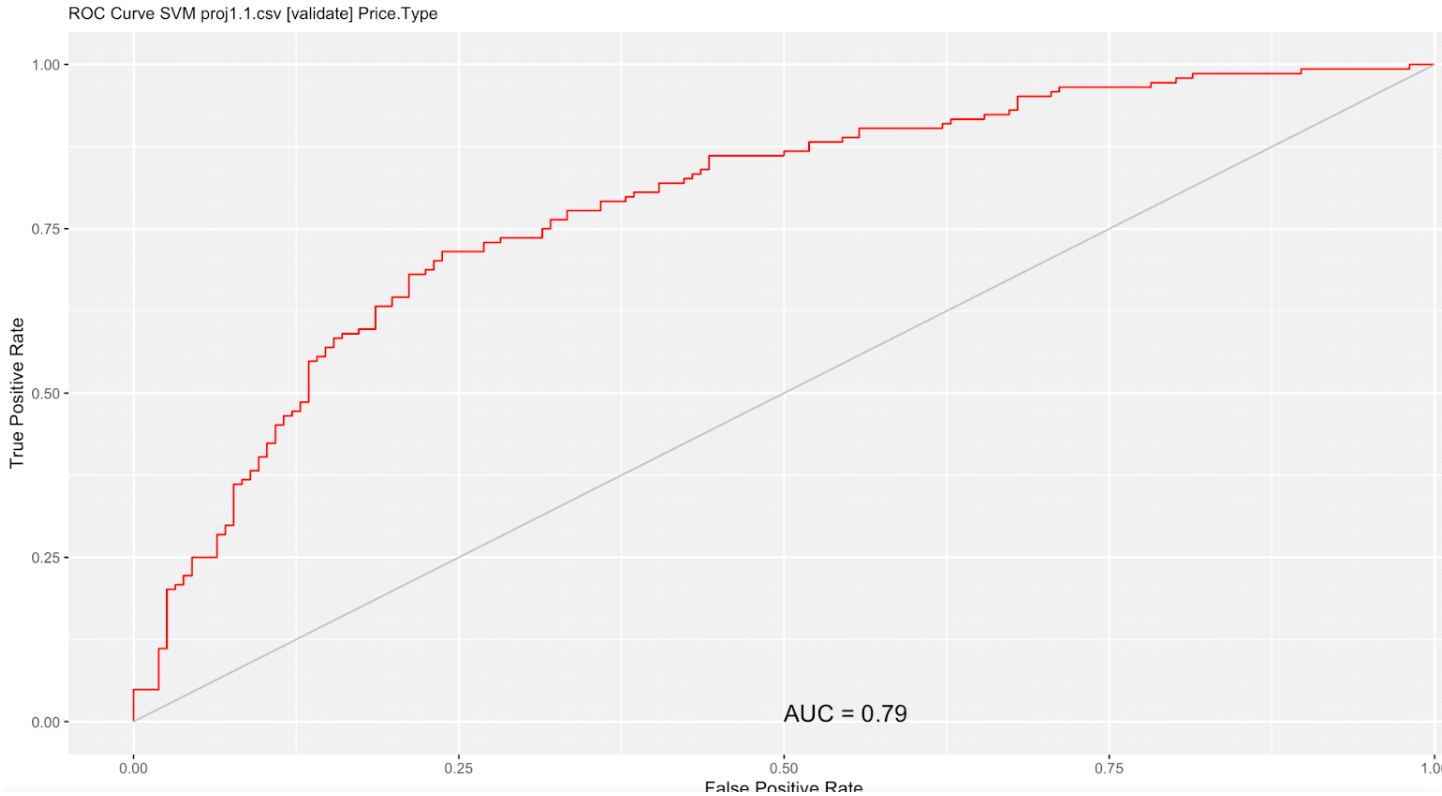
One large limitation of our dataset is that it doesn’t include data on technological features built into the various cars. If we had information on if the car has a GPS system built-in or Bluetooth compatibility that would be another factor influencing the price of the vehicle. Information regarding whether the car has heated or air-conditioned seats and how many would also be an interesting feature to look at while determining how it relates to the price of the car.

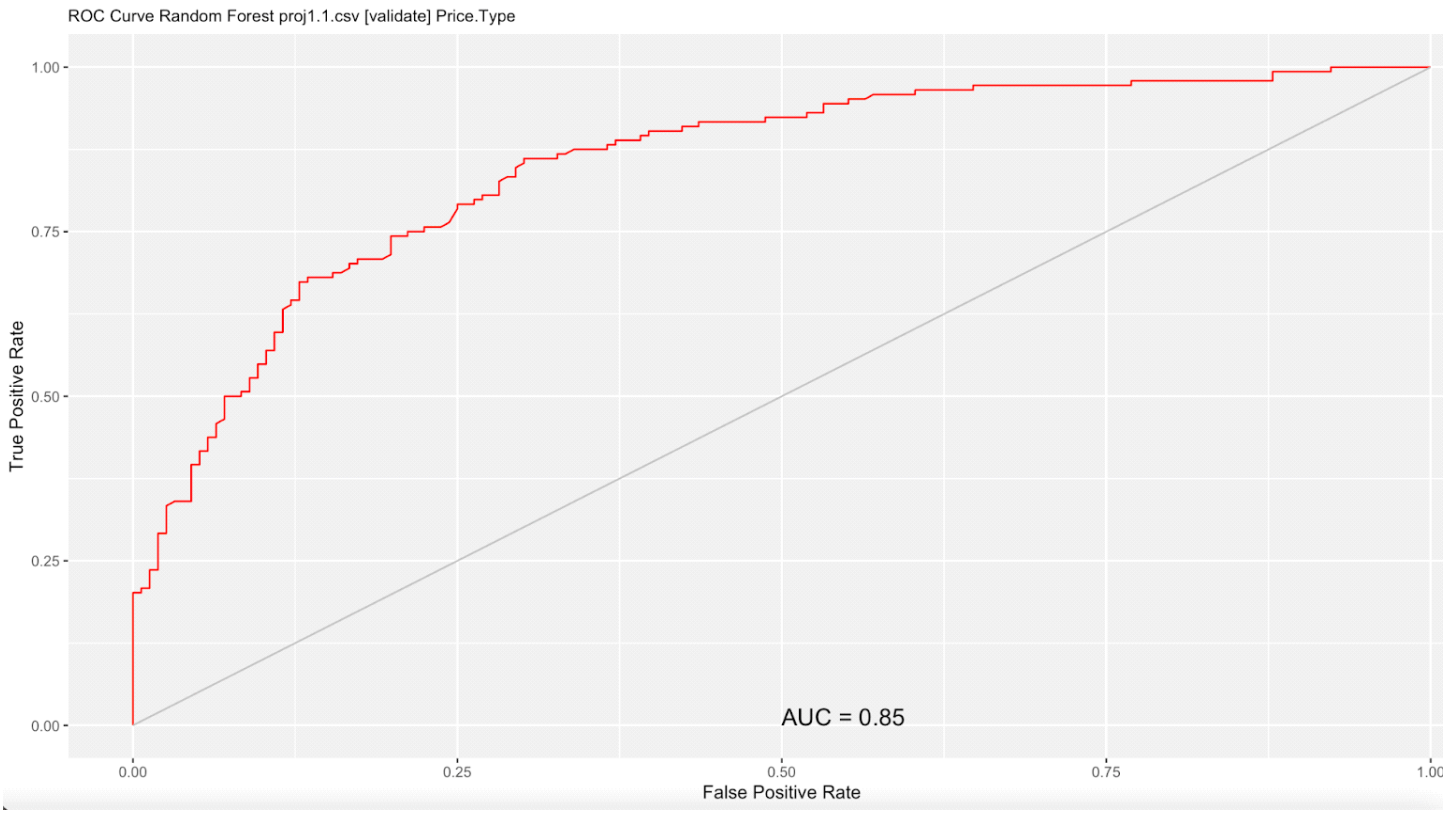
Another limitation of the dataset is that it doesn’t have information regarding the car’s miles per gallon. This could be an important factor in determining the price of a car since consumers might be willing to pay more for a car that gets more miles to the gallon. If we had this information, we could make visualizations and analyze how close of an effect the miles per gallon has on the overall price of the car.

Lastly, the third limitation of our dataset is that it does not contain information on if the car has a backup camera. As technology advances, some of the newer used cars in the dataset may have this feature while older models may not. This could influence the value of a car making it more expensive than others without one since it is a feature that exists within many vehicles now.

Appendix


ROC Curve Decision Tree proj1.1.csv [validate] Price.Type
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC = 0.71
0.00
0.25
0.50
0.75
1.00
False Positive Rate


ROC Curve SVM proj1.1.csv [validate] Price.Type
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC = 0.79
0.00
0.25
0.50
0.75
False Positive Rate




```
Variable Importance
====================

                                  Cheap Expensive MeanDecreaseAccura
R01_Airbags                       22.89     25.00                30.
TIN_Wheel_Right.hand.drive        22.01     25.80                28.
TNM_Levy                          25.98      6.34                26.
TNM_Engine.volume                 12.89     20.51                25.
TIN_Gear.box.type_Tiptronic       14.57     24.18                25.
TIN_Manufacturer_HYUNDAI          17.43     12.27                21.
TIN_Leather.interior_Yes          17.34     12.16                19.
TIN_Category_Jeep                 12.96     14.47                18.
TIN_Gear.box.type_Manual          12.70     13.46                17.
TNM_Mileage                       15.42      8.38                17.
```

https://www.cnbc.com/2022/07/21/consumers-paying-average-10000-above-normal-prices-for-used-cars.html